CMOS Devices and Beyond – A Process Integration Perspective


*Semiconductor Research Corporation, Research Triangle Park, NC 27709, USA
**Dept. of Material Science, North Carolina State University, Raleigh, NC 27601, USA
*** Dept. of External Programs, Intel Corporation, Austin TX 78746

Abstract. Development of CMOS technology is approaching severe technological limits in the next 10 – 15 years. Overcoming these limits will demand introduction of new manufacturable materials and device structures to extend the speed of silicon integrated circuits at the historical rate of 17 % per year to the end of the 2001 International Technology Roadmap for Semiconductors (2016). Following a brief discussion of these limits, this paper will review the most promising approaches to new materials, device structures and issues related to their integration in advanced CMOS structures. The paper will conclude with some brief observations and issues regarding extension of CMOS-like FET structures via new nano-scale materials.

INTRODUCTION

The accelerating pace of scaling CMOS technology, as seen in the 2001 Semiconductor Industry Association’s International Technology Roadmap for Semiconductors (ITRS), is quickening introduction of a variety of new materials, process technologies and transistor structures [1]. Some combination of these innovations undoubtedly will be required to sustain the historical increase of transistor speed at 17% per year for high performance logic, while simultaneously controlling exponentially increasing leakage currents and power dissipation critical for low power mobile applications. While silicon MOSFETs may approach their physical limit of operation at physical gate lengths of 9-nm within the next 10 to 15 years, intense pressure will continue to obtain increased cost and performance benefits from information processing technologies beyond the scaling of CMOS.

In response to the notion that the end of CMOS scaling is in sight, the microelectronics research community is accelerating exploration of a variety of new schemes for information processing. These approaches range from using new binary logic devices realized using molecules, nanowires, nanotubes, etc., to replace MOSFET switches to completely new algorithms for information processing, e.g., quantum computers or quantum cellular automata.

The goal of this paper is twofold. First we review new materials, process technologies and device structures being pursued to extend CMOS technology to and beyond the end of the 2001 ITRS. Then we address two specific issues related to new information processing technologies. One issue relates to comparison of the highly diverse new paradigms proposed for information processing. To address this need, a set of criteria by which new information processing technologies might be compared is proposed. The second issue regards the fundamental limit of scaling of charge-based logic switching devices, based on first principles. To address this issue, we offer an analytical commentary on the application of low-dimensional structures (molecules, nanowires and nanotubes) to MOSFET-like logic devices embedded in a silicon process architecture.

CMOS SCALING

Driven by Moore’s Law [2] of quadrupling the device density every 2 – 3 years, scaling of CMOS transistors for the past 30 years has followed a rather consistent trend guided by a set of scaling rules. These rules were first described by Dennard [3] to sustain constant electric fields in the gate oxide and the channel, and, thereby, to ward off the plague of the short channel effect (SCE). This relentless scaling is now entering the realm where several structural features of the MOSFET are beginning to reach atomic dimensions. For example, thickness scaling of the gate oxide to a few atomic layers has reached the point where the exponentially increasing gate tunneling current is beginning to compete with the channel leakage current as being a dominant source of off-state power dissipation [4, 5]. Another example is found in the polysilicon gate electrode, where the doping concentrations are reaching the limits of solid solubility and the polysilicon depletion capacitance is an appreciable fraction of the total gate capacitance [6]. Other examples are found with the increasing resistance of the source and drain contacts, the increasing effect on threshold voltage variance caused...
by statistical fluctuations of channel dopants, and others [1].

Each of these emerging challenges is driving one or more possible solutions based either on introduction of new materials and processes, major new device structures, or, most likely, a combination of both approaches. We outline these changes below and point out potential needs for new metrologies and characterization methodologies to deal with these new approaches.

New Materials

A preferable means of scaling CMOS to the end of the roadmap would be to introduce manageable changes of materials and processes to the current bulk CMOS structure. This approach will likely be viable down to the 65-nm and even, perhaps, the 45-nm nodes and beyond. However, many believe that technologies addressing the 45-nm, and, certainly, the 32- and 22-nm nodes will require significantly altered device structures as well as changes in the materials and processes. This section discusses material changes. Changes in device structures are discussed in the next section.

Shrinking vertical dimensions are reducing the gate dielectric (silicon-oxynitride) thickness down to 1-nm or 4 – 5 atomic layers. At this thickness the gate tunnel leakage current is becoming prohibitively large and is approaching being comparable to the channel leakage current [4]. One solution is to replace the current gate dielectric with a new high-dielectric-constant material, such as hafnium oxide, hafnium silicate, yttrium oxide etc. [7]. This would allow the industry to reduce gate leakage currents by several orders of magnitude for the same gate capacitance (i.e., thicker gate dielectric) or to continue scaling of gate dielectrics down to oxide equivalent thickness of 0.5-nm. However, introduction of high-K dielectrics for the gate insulator still faces many materials and integration challenges, the solution of which, will decide which material is used. These challenges include simultaneously obtaining the correct value for the dielectric constant (K ~ 20) in a material that is stable when subjected to CMOS processing temperatures. Furthermore, this material cannot chemically react with either the gate electrode material or with the underlying silicon channel and it must not significantly reduce the mobility of the channel carriers.

Dimensional scaling of MOSFETs also is drawing more attention to the source/drain [8,9] and the gate electrode technology, which currently is doped polysilicon (n+ for n-channel and p+ for p-channel). Three issues currently impacting doped polysilicon gate electrodes for bulk CMOS [10] are 1) polysilicon depletion capacitance is becoming comparable to the gate capacitance, 2) for boron-doped, p+ polysilicon boron penetration into the gate dielectric and into the channel, and 3) increasing resistance of the gate electrode. A metal gate electrode technology that provides a mid-silicon gap work function or, ideally, a binary metal alloy that provides an adjustable work function governed by the alloy composition is an attractive approach [10]. Advanced structures such as ultra-thin body fully depleted SOI (UTB-SOI) and double-gate MOSFETs, discussed below, also may require an adjustable work function binary metal alloy gate technology. This is needed to provide symmetric threshold voltages for the n- and p-channel of a proper value, given that we would not be able to provide threshold adjustments by appropriately doping the MOSFET channels. The notion of not doping the channels of new MOSFET structures is attractive for several reasons. Using undoped channels will relieve challenges related to integration of the dopant processes, will dramatically reduce the problem of statistical variations of channel dopant on threshold voltage fluctuations and will provide for significantly improved channel mobility and on current, $I_{on}$.

Scaling of the MOSFET has enabled modest increases of $I_{on}$ over the past several years. Another opportunity to sustain or perhaps increase $I_{on}$ in future technology nodes is to use an ultra-thin layer of strained silicon grown on top of relaxed silicon-germanium layer in the channel region. This approach, referred to as strained silicon epitaxy, can increase the mobility of n-type material by as much as 60% and p-type material up to 30% [11]. (A related approach is to use germanium in the channel to obtain higher $I_{on}$ and more equally sized n- and p-channel MOSFETs [11a].) At least two companies, IBM and Intel, have announced plans to incorporate strained silicon channels in their future process technologies, as early as the 90-nm node. Integration of strained silicon, however, into some of the new device structures, particularly double gate structures including FinFETs discussed in the next section is expected to be very difficult.

New Device Structures

While introduction of high-K materials to deal with MOSFET scaling challenges is promising, new non-classical device structures also have been demonstrated to deal quite effectively with those challenges related to electrostatic field scaling. For example, short channel effects, dominant at the most advanced technology nodes at and beyond the 45-nm node, are substantially reduced by fully depleted SOI and by double gate structures. Two specific examples of these MOSFET structures are the ultra-thin body
fully depleted SOI (with a body thickness ≤ 10-nm) [12 - 17] and the FinFET [18 – 22]. Schematic illustrations of these two structures are shown in Figures 1 and 2.

The ultra-thin body fully depleted SOI MOSFET offers several advantages related to scaling CMOS to and below the 45-nm node, principally by reducing the short channel effect. The body thickness of this structure must be less than 1/3 of the gate length, L_g. One advantage of this structure is the channel height, separating the substrate isolating insulator from the channel – gate dielectric interface, is controlled by body thickness and not by the channel doping concentration. This provides vertical control of the channel current, and, to some extent thereby, of the short channel effect. Consequently, the body of the UTB-SOI device does not need to be heavily doped to control the short channel effect. This provides the additional advantages of higher channel mobility and, therefore, higher saturated drain current, I_{sat}, and much lower fluctuations of the threshold voltage, V_{th}, caused by statistical fluctuations in the channel doping density. However, this approach requires that adjustment of V_{th} for n- and p-channel MOSFETs to be obtained by adjustment of the work function of the gate electrodes, e.g., using a binary metal gate with a continuously adjustable alloy composition to obtain an adjustable work function of the gate electrode. Compared to bulk MOSFET structures, the UTB-SOI also allows more aggressive scaling of the gate dielectric thickness to sustain the gate leakage current. This is due to a lower vertical electric field in the UTB-SOI structure. Another attribute of this structure is that formation of ultra-shallow source/drain junctions is not required, since use of raised source/drain contacts is required to achieve the very low source/drain resistance [22].

Fabrication of UTB-SOI introduces many challenges; the first being formation of a uniformly thin silicon layer with thickness less than 10-nm to form the fully depleted body. Other challenges include doping of the thin silicon layer, selective Si_{1-x}Ge_x growth on the ultra-thin silicon layer necessary to form the required ultra-low resistance raised source/drain contacts and introduction of a relaxed Si_{1-y}Ge_y layer followed by growth of a strained silicon layer for enhancement of channel mobility. Lastly, a Schottky barrier source/drain contact technology may be needed for body thicknesses less than 5-nm to obtain ultra low source/drain resistance [23].

Several variants of the FinFET or a multiple-gate FET also are candidates either to succeed the UTB-SOI at the 32- to 22-nm node or, perhaps, to completely supercede the UTB-SOI structure beginning at the 45- to 22-nm node. Double gate structures, including the FinFET, relax the required body thickness to be less than 2/3 of L_g, compared to the UTB-SOI single gate MOSFETs as described above. Different structures belonging to the FinFET class of devices are principally distinguished by how the gate defines the channel(s) in the fin and by the aspect ratio (height/thickness) of the fin. In one structure, the gate surrounding the “dorsal-like” fin defines a channel only on the two vertical surfaces facing one another. These two gates can be electrically connected to form identical channels (double gate) or they can be separated to provide a fourth terminal (split gate), one terminal each for the two gates. In either case, the gate electrode crossing the top of the fin is completely isolated from the top of the silicon fin. In another variant, referred to as the Omega, Ω, or Trigate transistor, the gate defines a continuous channel on the fin’s top surface as well as the two vertical surfaces. In this case the gate forms a single gate electrode and the aspect ratio of the fin is...
somewhat lower than that for either the double gate or the split gate FinFET structures.

The FinFET (as well as other double gate MOSFETs) offers several advantages compared to bulk MOSFETs and to UTB-SOI structures. Chief among these are reduced short channel effects (i.e., more scaleable) and higher drive current, $I_{on}$. The reduced short channel effects result from superior thickness control of the channel current and better confinement of the drain induced electric field, by termination of the field lines on both gates located on the two sides of the channel. Enhanced $I_{on}$ results from steeper subthreshold voltage swing (approaching the theoretical limit of 60 mV/decade) and lower channel electric field resulting in higher carrier mobilities.

Double gate MOSFETs, particularly the several types of FinFET structures, face many barriers to their adoption as a manufacturable CMOS technology. First a sub-lithographic process will likely be needed for FinFETs to fabricate the fins with an extremely thin thickness dimension < 10-nm. Further, a very high selectivity gate etch will be required to etch the gate on top of the fin, which poses a constraint on the integration of high-K dielectric and metal gate electrode. A uniformly high source/drain doping through the fin depth is required to obtain high $I_{on}$ and providing a strained silicon layer on the vertical faces of the fin to enhance carrier mobility will be very challenging. An additional issue is the difficulty of fabricating and preserving the high-aspect-ratio fins. One approach recently proposed [22] for dealing with this issue is to use the Ω or the Trigate FinFET structure. This structure allows use of a lower aspect-ratio fin structure due to its ability to provide control of the charge in the fin by controlling the gate-channel interface potential on three surfaces of the fin. Another issue impacting design methodologies and design tools is the discrete nature of the gate width of the FinFET structure. The gate width of the FinFET is set by the fin height, which is fixed by the process and, therefore, is not continuously adjustable. For this reason, total transistor current is obtained only in discrete incremental steps determined by the number of fins per transistor. Also, the Ω or the Trigate FinFET structure being three –terminal device appears to be limited in its ability to provide dynamic control of $V_{th}$.

**BEYOND CMOS SCALING – FUNDAMENTAL LIMITS TO TRADITIONAL SCALING**

As the semiconductor industry approaches the scaling limits of silicon CMOS devices and integrated circuits, the research community has begun searching for alternative devices and information processing technologies to sustain the traditional device scaling and, perhaps, to evolve to a new paradigm of “functional” scaling. New device approaches that are being considered for alternative charge-based switches include molecular and carbon nanotube structures, single electron transistors, etc. New information processing paradigms currently being explored include Quantum Cellular Automata, Complex Nonlinear Networks, Quantum Computing and Quantum Associative Memories. With such a rich diversity of approaches pursuing multiple fundamentally different paths of exploration, the need for their critical evaluation on some basis becomes important. The intent is not to make very premature choices between alternative competing approaches, but to gain insights of the potential/risk tradeoffs among the leading candidates ideally using a common set of metrics. A set of metrics proposed by the ITRS’s Emerging Research Device Working Group is discussed below.

**Proposed Metrics for New Information Processing Technologies**

A set of metrics proposed by the 2003 ITRS’s Emerging Research Device Working Group is outlined below. These metrics are divided into metrics for “near-term”, circa 2011 – 2020, and “long-term” beyond 2020.

**Near-Term Metrics (2011 – 2020)**

Metrics proposed for the “Near Term” should address technologies that will enhance and extend silicon CMOS such that they can be in high volume manufacturing in the period of 2011 to 2020. These metrics include:

1. Extend the performance of microelectronics significantly beyond the 22-nm node while simultaneously being cost competitive with CMOS on a per function basis,
2. Provide methodologies for connecting individual information processing elements and storing information/data,
3. Be stable on the scale of a few atoms under a wide range of processing and operating environmental conditions,
4. Operate at or near room temperature,
5. Provide an energy restorative process(es) within the technology infrastructure for irreversible systems (gain or signal level restoration),
6. Be capable of physical realization in the presence of statistical variations of process and performance parameters, and
7. Be compatible with CMOS manufacturing processes.

Long-Term Metrics (Beyond 2020)

Metrics proposed for the “Long Term” should allow for technologies that are based on new physical mechanisms and which are likely to obey completely new scaling laws. We would hope that a technology that extends microelectronics in the Near Term might provide a platform to launch a new approach addressing the Long-Term needs. Metrics addressing Long-Term approaches to information processing include:

1. Provide potential advantages orders of magnitude beyond the technology performance metrics (size, density, speed or information throughput rate, power dissipation, etc.) compared to the 22-nm ITRS node for CMOS,
2. Satisfy metric numbers 2, 3, 4, 5 and 6 in the Near Term list above,
3. Scaleable over multiple generations with an exponential improvement in technology performance metrics for an incremental increase of cost,
4. Provide a reasonable balance between technology potential and risk, and
5. Be compatible at the systems level with CMOS infrastructure.

New approaches to information processing technologies are exploring many new mechanisms to represent and process, store and communicate information represented as bits. These mechanisms range from using new, ultra-scaled structures (e.g., molecules, nanotubes, nanowires, etc.) to replicate current means for representing bits using charge, voltage or current to use of “entangled” quantum states represented by the state of a qbit (or quantum bit). Discussion of this broad range of topics is beyond the scope of this paper. However, one important question related to new charge-based information processing approaches addresses the fundamental limits (size, energy, speed, etc.) of an elemental switch. The next section briefly addresses this question.

Fundamental Limits to Scaling of Charge-Based Nanoelectronic Switch Elements

An important issue regarding charge-based nanoelectronic switch elements is related to the fundamental limits to the scaling of these switches, and how these new devices compare with CMOS at its projected end of scaling. The 2001 ITRS projects the scaling of CMOS down to the 22-nm node, which represents a physical gate length for an MPU/ASIC device of 9-nm with a power dissipation of 93 W/cm². An analysis summarized in this section concludes that the fundamental scaling limit of a charge-based “gedanken” switch is only a factor of 5 – 10 smaller than the physical gate length of an end of the ITRS CMOS device. Further we find that the density of these switches is limited by maximum allowable power dissipation of approximately 100 W/cm², and not by their size. A more complete analysis was presented at the recent Trends in Nanotechnology Conference [24] in Santiago de Compostela, Spain, and in the IEEE Proceedings [25].

A “gedanken” switch can be represented as two potential wells of width, \( \omega \), separated by a rectangular energy barrier of width, \( \alpha \), and energy height, \( E_b \), as shown in Figure 3. Further, we will define \( E_b \) to be the energy difference between the lowest allowed energy state in the well and the top of the energy barrier.

![Fig. 3: Schematic diagram of the “Gedanken” switch consisting of 2 potential wells separated by a biased finite barrier with one of the wells containing an electron the position of which defines the logic state.](image)
The total linear dimension of this switch is given by \( \lambda = 2a + \alpha \). Imagine that one well contains a charge at a lowest level energy state and the other well is empty. Sustaining the charge in its state in the chosen well is governed by the energy barrier separating the two wells, particularly by the width and height of the barrier, which together control charge transfer from one well to the other. The maximum density of these "gedanken" switches is given by \( n = \frac{1}{(\alpha^2)} \) and the power dissipated in changing the state of this switch (i.e., moving the charge from one potential well to the other through lowering the barrier, \( E_b \)) is given by \( P = \frac{E_b}{\tau} \). Here \( \tau \) is the time required to transfer the charge or the switching time. The minimum switching energy per operation for a reversible computation is given by the Shannon-von Neumann-Landauer (SNL) limit that is given by \( E_b = KT \ln 2 \), or 0.017 eV for \( T = 300 \) K. Using Heisenburg’s Uncertainty relationships, one determines that for minimum energy of operation the minimum size of this switch \( \lambda \) is approximately 1.5 nm. Similarly, the minimum switching time is determined to be \( \tau = 0.04 \) ps and the maximum switch density is approximately \( n = 5 \times 10^{15} \) devices/cm\(^2\). The maximum power density \( P_d \) (assuming only the density given above) is \( P_d = nP = 5 \times 10^6 \) W/cm\(^2\), which is a factor of \( 5 \times 10^4 \) larger than our allowed power dissipation of 100 W/cm\(^2\). To achieve this limit power dissipation, we would need to lower the device density from \( n = 5 \times 10^{15} \) devices/cm\(^2\) to \( n = 1 \times 10^9 \) devices/cm\(^2\), which is approximately the density of the end of the ITRS CMOS.

Based on this simple analysis, we conclude that:

1. Any charge-based switching technology will be limited in its density by the maximum allowed power dissipation and not by the limit size of the switching element.
2. The limit size of any charge-based switch is only approximately a factor of 5x smaller than the minimum physical gate length of a silicon-based MOSFET transistor at the end of the ITRS.
3. The end of the ITRS CMOS MOSFET technology potentially could be nearly an ideal switch that is based on charge.
4. Finally, we think that research on alternate charge-based switch technologies might best be focussed on replacing the channel with an alternate material in an otherwise standard silicon MOSFET fabrication infrastructure.

**SUMMARY AND CONCLUSIONS**

Development of silicon-based CMOS technology is approaching severe technological limits in the next 10 – 15 years. Overcoming these limits will demand introduction of new manufacturable materials and device structures to extend the speed of silicon integrated circuits at the historical rate of 17 % per year to the end of the 2001 International Technology Roadmap for Semiconductors. Following a brief discussion of these limits, this paper reviewed the most promising approaches to new materials, device structures and issues related to their integration in advanced CMOS structures.

Finally, we conclude that silicon-based CMOS technology scaled to the end of the 2001 ITRS will likely yield a near-ideal charge-based switch and that at this limit the device density will be limited by power dissipation and not by device size.

**REFERENCES**


